Referee report on 'A Replication of "The Role of Intermediaries in Facilitating Trade" (Journal of International Economics, 2011)' submitted to *Economics: The Open Access, Open Assessment E-Journal* (MS 3353)

This paper replicates Ahn, Khandelwal and Wei's (2011) results for three testable predictions of a model examining intermediaries' role in exports using Chinese firm-level data for 2005. Using independently collected data, AKW's results are confirmed when the same models are fitted. Additional tests are then conducted including (i) splitting the data for China into three geographical subsamples (East, Central, West); (ii) using customs data for separate years from 2000 to 2004; (iii) using 2012 Enterprise Survey data (that post-dates the original study); and (iv) using alternative versions of the World Bank's 'Doing Business' data on the number of documents required for import procedures in China. These additional estimations provide a much more mixed set of results in terms of their support for AKW's predictions.

The paper is well-written and the various steps in the replication analysis are clearly explained. It represents a good example of a replication analysis that engages with and achieves cooperation from the authors of the original study. It also makes a compelling case that there is no incentive to 'cherry-pick' results that confirm or refute those in the original study. This is a useful exercise and there are no obvious concerns about the implementation of the analysis. The main points below relate primarily to the interpretation of the results in terms of how far they support or undermine AKW's predictions.

Main points
1. The replication analysis follows the original AKW study in examining three measures of 'productivity' – sales, employment and sales per worker – in the models for direct and indirect export shares in Table 1. The original results are successfully reproduced. For direct export shares, a statistically significant (at the 5% level) positive linear effect is obtained only for the sales per worker measure. For indirect export shares, the quadratic term is not significant (at the 5% level) for the sales per worker measure, but is significant for sales and employment. Sales per worker is a standard measure of (labour) productivity, but sales and employment are really measures of 'size', not productivity. The differences in the results for the different measures are therefore not entirely surprising. Prediction #1 is explicitly concerned with the effects of changes in productivity not size. In the absence of evidence that sales and employment are reasonably well correlated with productivity, it makes sense to put more weight on the results for labour productivity in the original results and in the extensions testing Prediction #1 in Table 4. The latter results suggest either that effects are very imprecisely estimated (for direct export share) or contradict the theoretically expected signs (for indirect export share). On this interpretation, the results in Table 4 are damaging to AKW's Prediction #1.

2. Considering regional sub-samples, as in Table 5, is a reasonable exercise because the model in Table 2 (columns (3.a) and (3.b)) implicitly assumes constant parameters for the relevant population of all intermediary firms. Given that the results for the East region, as with the full-sample results, are consistent with Prediction #2 but those for the Central and West regions are not, the main issue is how to interpret these differences. Economic activity is more concentrated in the coastal East region, which, consequently, is much more highly represented in the overall sample. Therefore, it is not surprising that the East sample results are similar to those in Table 2. Other factors are likely to be relevant for the interior regions (beyond product, ownership and destination country fixed effects), but the theoretical model is not intended to consider sub-national geographical factors and the minimal empirical results reported provide no guidance as to whether the model specification is appropriate for

these regions. It is hard to judge whether the results for Central and West are to be considered as evidence against the mechanisms in the AKW model or are the product of other types of model misspecification. The latter may also be a possible explanation of the volatility in the estimates for the different years. In addition, it is recognized that the 2005 data are preferable because of the absence of license restrictions. Overall, the implications for Prediction #2 of the differences in the regional and year-by-year estimates are harder to interpret, although the instability in the estimates does cast doubt on any results from the empirical model.

3. For the robustness checks for Prediction #3:
(a) The discussion of regional results in 2 is again relevant. In particular, the theoretical model predicts that higher variable trade costs and higher fixed costs of exporting should affect intermediaries' share of trade. These costs are measured by home- or destination-country proxies at the national level; any additional regional cost factors are therefore not considered. These could be related to additional transportation costs given the vast internal distances involved, distance to the coast, etc. It is not obvious what the effect of omission of such variables would be but it could contribute to variation in regional estimates.
(b) Some of the proxies are likely to be more relevant than others. Indeed, as noted, the results for distance and GDP are consistent across all model variants, in line with AKW's Prediction #3. In addition, the online measure of the number of import documents is held to be the preferred measure and this provides results more supportive of the AKW prediction than the other variants of this variable. Overall, taking into account the various difficulties of measurement and interpretation, it seems to me the results are not as unsupportive of Prediction #3 as an equally weighted 'success rate' of statistically significant estimates with expected signs might suggest.

Minor points
p.1, bottom line: Add 'of' after 'in favor'.
p.3, line 5: Tang and Zhang (2012) is not in the reference list.
p.3, last para, line 1: Insert grave accent over the 'a' in 'Antras'.
p.7, para 2: There appears to be a contradiction, or at least some ambiguity, between stating that "while responses are collected at the firm-level, they are only available to researchers at the country-level" and saying that the survey provides proportions of individual firms' sales exported directly and indirectly, as well as various productivity measures. Clearly, the latter is correct.
p.9, para 3, line 5: 'large' rather than 'great'? Related to this, given that the sample sizes are often very large, why persist with using conventional 10%, 5% and 1% significance levels to indicate statistical significance, rather than adjusting the significance level to take into account the sample size?
p.9 and elsewhere in the main text: US spelling is used for 'gray', but UK spelling, 'grey', is used in the table notes.
p.12: In the robustness check for prediction #2 that splits the data by geographical region, only the full model's results (corresponding to column 3a in Table 2) are reported. A desire to avoid proliferation of results is appreciated, but do the results for the other (simpler) models yield similar results to those in Table 5? (As it stands, there is no obvious way of selecting a 'best' model from Table 2, as only a goodness of fit measure is reported, and this is bound to favour model 3 because of the additional regressors.)
p.14, fn. 4: Add full stop and closing quote marks.
p.20, Table 5: The 2000 result is shaded red, but the estimated coefficient, though having the wrong sign, is not statistically significant at the 5% level. For consistency, this cell should be rose-tinted. Similarly, the statement in the text that "of the remaining three, one is wrong-signed and insignificant (2004), and two are wrong-signed and significant at the 1 percent

level (2000 and 2003)" treats the 2000 result as significant at the 5% level. Is the standard error for this result correctly reported?

p.20, last para: It is claimed that this paper applies Mayo's 'severe testing' approach. But misspecification testing plays a big part in her approach, which is not considered either in the original study or the replication exercises. Therefore, although several useful dimensions of the modelling are probed, it is unlikely that Mayo would consider these sufficient to be considered a severe test.

p.21, para 3: "… between productivity and labor productivity" – this should refer to productivity and indirect exports.

p.21: Finishing on a positive note for future research is good!

p.36, Figure 1: is there a need to define the parameters in the model that appear in the slope term expressions? Either the reader could be referred to AKW's paper or the slope term expressions could be removed.